RESEARCH PAPER

Multi-Output Gradient Boosting for Interpretable Multi-Disease Prediction Using Minimal-Feature Survey Data

Sajja Tulasi Krishna[1], K.R. Manjula Devi[2]
[1,2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, AP
Email: tulasisajja788@gmail.com
Email: manjuladevikr.2511@gmail.com

Received: 1st Sep, 2026 | Revised: 8th Sep, 2026 | Accepted: 15th Sep, 2026 | Available Online: 22nd Sep, 2026

ABSTRACT: This research is about Chronic diseases including diabetes mellitus, cardiovascular disease (CVD) and chronic kidney disease (CKD) have strong pathophysiological interdependencies, but most machine learning predictors either consider them separately or require longitudinal electronic health records unavailable in resource-constrained settings. This work helps fill an important gap by presenting some lightweight and interpretable multi-disease frameworks that is works for population-level screening based on minimal feature survey data. We developed a multi-output gradient boosting classifier trained with 70692 respondents from the Behavioral Risk Factor Surveillance System (BRFSS) using only 19 demographic and lifestyle features, SHapley Additive exPlanations (SHAP) for model interpretability and rigorous external validation on an independent cohort. The framework was able to simultaneously predict diabetes (area under the receiver operating characteristic curve [AUC] = 0.831, 95% CI: 0.826-0.837) and CVD (AUC = 0.809, 95% CI: 0.803-0.817) and CKD risk proxy (AUC = 0.962, 95% CI: 0.960-0.964) agents with an exceptional stability during external validation (|ΔAUC| < 0.0 SHAP analysis resulted in rules clinician friendly, such as 'Age under 65 years increases CKD risk by +1.262 SHAP units, which were confirmed through structured physician's feedback forms. The results show the practicality of implementing clinically relevant multi-disease prediction based on resource-efficient ensemble techniques, independent of EHR systems and GPU-intensive models, which is crucial for the deployment of such approaches in primary care settings where longitudinal health records are very rare to access.

Keywords: Chronic disease prediction; multi-task learning; Gradient boosting; Explainable AI; Population health screening; Behavioral Risk Factor Surveillance System; Diabetes mellitus; cardiovascular disease; chronic kidney disease; Clinical decision support

How to cite this article: Sajja Tulasi Krishna1 K.R. Manjula Devi2. Multi-Output Gradient Boosting for Interpretable Multi-Disease Prediction Using Minimal-Feature Survey Data. Int J Drug Deliv Technol. 2026;16(80s):674-683. DOI: 10.25258/ijddt.16.80s.83.

INTRODUCTION

Chronic non-communicable diseases, such as diabetes mellitus, cardiovascular disease (CVD) and chronic kidney disease (CKD) are the leading causes of death worldwide, representing unsustainable burdens for healthcare systems worldwide [4]. Critically, there are very strong pathophysiological interdependencies involved: diabetes patients have a substantially high risk for CKD [4], and the presence of CKD is an independent risk factor for CVD mortality [22]. Despite this clinical reality, traditional methods of machine learning generally use single disease prediction models that do not account for patterns of comorbidity, and thus, fail to harness common risk pathways for better diagnostic accuracy [12].

Recent studies are towards multi-task learning (MTL) architectures are jointly developed for model disease correlations. Kim et al. [14] demonstrated how intra-person MTL frameworks can be constructed on CNN-LSTM networks to promote the stability in predicting diabetes hypertension using the temporal feature representation sharing across the related tasks. Similarly, Zhang et al. [28] created MTL-Cox models with substantial accuracy improvements compared to single-task baselines for nine chronic conditions on UK Biobank data. Men et al. [21] extended LSTM networks with time-aware and attention mechanisms for multi-disease prediction with more than five million clinical records. However, these complex frameworks demonstrate a critical dependency on longitudinal electronic health records (EHRs) with sequential clinical visits a resource unavailable in the majority of low- and middle-income countries according to systematic reviews [4]. This dependence on EHR data represents a translational hurdle: While the field is now showing promise on a theoretical level, its clinical value has been limited to settings that are well resourced and have advanced health informatics systems. At the same time, the field has witnessed an explosion towards increasingly computationally demanding architectures.

Concurrently the field is more inclined to more computationally intensive model architectures. Dao et al. [3] recently released CURENet a multimodal model combining LLMs for clinical notes, transformers for time series, and structured labs with 94% accuracy on 10 chronic conditions. Tsai et al. [24] similarly proposed attention-based multimodal

networks which capture cross-disease dependencies. However, these methods necessitate GPU intensive inference and structured EHR ecosystems and, thus, are not amenable for population level screening in resource limited primary care settings with only basic demographic and lifestyle information available. [4]

This dichotomy highlights a core tension in predicting chronic disease research: accuracy vs. accessibility. Deep learning and LLM-based systems can suffer small performance improvements at prohibitive computational costs and lightweight ensemble systems are limited to single disease prediction despite their deployment benefits. The systematic review of Delpino et al. [4] pointed out critically that less than 15% of chronic disease prediction research employs nationally representative survey data such as BRFSS describing a large gap in frameworks that are applicable to screening cross sectionally in the population. Yang et al. [26] focused on predicting at the population level using gradient boosting on 451,425 working adults (AUC=0.850) but only predicted aggregate "multiple chronic conditions" without disease-specific prediction of risk and/or external validation.

Furthermore, clinical adoption requires actionable interpretability beyond SHAP values in post-hoc. While the work by Khan et al. [13] presented explainable surveillance systems with physician validation, their frameworks still depend on EHR-derived features (labs, medications) that are unavailable in minimal resources screening settings. Ge et al. [7] came up with a CNN equipped with correlation-aware loss for multi-label prediction but at the cost of interpretability loss. The lack of clinician-interpretable rules based on survey-based risk factors is a critical issue in the deployment to the field [12].

To fill these gaps, we propose a multi output gradient boosting approach that predicts diabetes, CVD, and CKD simultaneously using 19 features from BRFSS nationally representative survey data only. This study is a lightweight survey based on multi-output predictor using combination of clinician rules derived from SHAP and externally validated stable performance ($|\Delta_f \text{AUC}| < 0.03$). By putting implementation feasibility over extra architectural complexity, the framework allows scalable multi-disease risk stratification in settings for which longitudinal EHR systems are lacking [4].

I. RELATED WORK:

Early applications of machine learning for chronic disease prediction were mostly single-disease classification predictors based on traditional ensemble methods. Islam [11] obtained 98.75% accuracy in chronic kidney disease (CKD) prediction by CatBoost with SHAP explainability, and other studies with Random Forest and AdaBoost on UCI datasets were reported to be able to return high accuracy for heart disease and by Random Forest [3]. However, these approaches considered diseases as individual entities in spite of well-established clinical comorbidities diabetes patients exhibit threefold higher CKD risk thus not taking advantage of shared pathophysiological pathways across related conditions.

In order to solve disease interdependencies, researches shifted to multi-label and multi-task learning paradigms. Ge et al. [7] proposed GroupNet, a convolutional neural network with correlation-aware loss functions to model the relationships between ten chronic diseases with an accuracy of 81.13%. Zhang et al. [29] and Maxwell et al. [20] further contributed to multi-label classification by problem transformation methods and deep architectures respectively. Nevertheless, these models were computationally intensive black boxes with limited clinical interpretability and were heavily dependent upon longitudinal data from electronic health records (EHR) which are not available in resource-constrained settings.

The most advanced frameworks had introduced some temporal modeling to model the dynamics of disease progression. Kim et al. [14] proposed an intra-person multi-task learning CNN-LSTM model that predicts both diabetes and hypertension jointly with the same LSTMs and demonstrated higher stability than single task baselines. Men et al. [21] extended Lstm network with time aware and attention mechanism for multi disease prediction using more than five million clinical records. Zhang et al. [28] built MTL-Cox, a survival analysis framework that predicts nine chronic diseases using UK Biobank data with 12% accuracy improvements compared to other competing methods. While effective, these approaches require sequential visits in clinical settings and structured EHR infrastructure resources which are not available in 78% of low- and middle-income countries according to systematic reviews [4].

Recent developments involve the use of multimodal data integration using large language models. Dao et al. [3] introduced CURENet, which integrates unstructured clinical notes, lab results, and time-series data using LLMs and transformers for the prediction of ten chronic conditions achieving 94% accuracy on MIMIC-III. Tsai et al. [24] used multimodal inputs with attention mechanisms in a similar way for improved interpretability. Despite the high predictive performance, LLM-dependent architectures are costly in terms of computational resources and structured EHR ecosystems, limiting their applicability to get deployed in settings of primary care where only basic demographic and lifestyle data can be found.

Explainability has become an important requirement for clinical applications. Khan et al. [13] created an explainable surveillance system based on SHAP values and rule engineering for Random Forest

models, which was validated by physicians based on routine EHR data. Although most of the frameworks offer post-hoc interpretability, they do not have clinician-actionable decision rules based on an established clinical threshold. Yang et al. [26] addressed population-level screening based on the gradient boosting algorithm with data from 451,425 working adults undergoing checkups (AUC=0.850) but did not predict specific disease risk or stratify individuals according to their specific disease risks and did not validate the results with independent data.

Delpino et al.'s systematic review [4], for example, critically noted that "less than 15% of studies on prediction of chronic disease used nationally representative survey data such as BRFSS" and therefore there is a significant gap in lightweight frameworks that can be used in cross-sectional population screening. Most of the existing multi-disease predictors either need longitudinal EHRs [14 21, 28], require GPU intensive LLM inference [3, 24], or lose specific disease interpretability to aggregate risk scores [26]. This context points to the need for resource-efficient multi-task predictors that can work on minimal feature sets and are accompanied by clinical transparency an important requirement for implementation in environments with limited health informatics infrastructure.

Contributions & Novelty Table

Contribution	Technical Implementation	Differentiation from Prior Work	Reference Gap Addressed
Multi-output ensemble for correlated diseases	MultiOutputClassifier(GradientBoosting) with shared feature space	First gradient boosting (not deep learning) for simultaneous diabetes/CVD/CKD prediction	[14] (CNN-LSTM MTL), [7] (GroupNet CNN)
Survey-data MTL framework	BRFSS 2022 dataset (19 features: demographics + lifestyle)	Only MTL predictor using nationally representative cross-sectional data (no EHR/time-series required)	[26] (aggregate MCC prediction), [4] (systematic review gap)
Clinician-interpretable rule extraction	SHAP global importance → binary decision rules → physician validation sheet	Rules grounded in clinical thresholds (e.g., GenHlth>3.0) + structured clinician feedback	[13] (SHAP only), [7] (attention maps)
Stable external validation	Train on BRFSS 2022 → test on BRFSS 2021 subset (19 shared features)	ΔAUC <0.01 across datasets rare in chronic disease ML literature	[7, 20, 29] (internal CV only)
Resource-efficient deployment	<500ms inference on CPU; no GPU/LLM dependency	Enables screening in primary care/LMICs where EHRs unavailable	[3] (LLM-heavy CUREN et), [24] (multimodal transformers)

II. METHODOLOGY

This study uses a multi-output gradient boosting approach for the joint prediction of diabetes, cardiovascular disease (CVD) and chronic kidney disease (CKD) with data from the nationally representative BRFSS survey. Our methodology combines disease correlation modeling using shared representations of features, pathophysiological cascade constraints are used in line with KDIGO/ACC guidelines, and SHAP-based explainability is used to produce clinician-actionable risk rules. The external validation on a separate cohort is important in order to ensure robustness without retraining.

3.1 Multi-Label Target Construction

We formulate chronic disease prediction as a multi-label classification problem where each patient may simultaneously exhibit multiple conditions. Let $X \in R^{n\times d}$ represent the feature matrix with n patients and $d = 19$ clinical attributes. The target space is defined as $Y = [y_{\text{diabetes}}, y_{\text{cvd}}, y_{\text{ckd}}] \in \{0,1\}^{n\times 3}$ where:

$$y_{\text{diabetes}}^{(i)} = \begin{cases} 1 & \text{if } Diabetes\_binary^{(i)} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\square_{\text{cvd}}^{(\square)} = \begin{cases} 1 & \text{if } HeartDiseaseorAttack^{(i)} = 1 \text{ or } St \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For chronic kidney disease (CKD) there is no direct field of diagnosis in BRFSS. We develop a probabilistic proxy based on known clinical risk factors based on KDIGO guidelines:

$$s_{\text{ckd}}^{(i)} = 3 \cdot 1\left(Age^{(i)} \geq 9\right) + 2 \cdot 1\left(\square\square\square h\square\square^{(\square)} = 1\right) + 3\square_{\text{diabetes}}^{(\square)} + 1\left(\square\square\square^{(\square)} \geq 30\right) + 1\left(Smoker^{(i)} = 1\right) \quad (3)$$

where $I(\cdot)$ denotes the indicator function. The continuous risk score is transformed to probability via sigmoid activation:

$$p_{\text{ckd}}^{(i)} = \frac{1}{1 + \exp\left(-\left(s_{\text{ckd}}^{(i)} - 5\right)\right)} \quad (4)$$

Binary labels are assigned with controlled noise injection to simulate diagnostic uncertainty:

$$\square_{\text{ckd}}^{(\square)} = \begin{cases} 1\left(p_{\text{ckd}}^{(i)} > 0.5\right) \oplus b^{(i)} \text{with probability } 0.05 \\ 1\left(p_{\text{ckd}}^{(i)} > 0.5\right) \text{otherwise} \end{cases} \quad (5)$$

where $b^{(i)} \sim \text{Bernoulli}(0.5)$ introduces 5% label noise and $\oplus$ denotes XOR operation.

CKD labels were made as a proxy of probability of risk based on known clinical risk factors (age, hypertension, diabetes, BMI, smoking) because of lack of laboratory confirmed eGFR or albuminuria measures in BRFSS. Therefore, CKD predictions reflect risk stratification as opposed to certainty of diagnosis.

3.2 Feature Preprocessing and Stratified Sampling

Numeric features undergo z-score normalization to ensure gradient-based optimization stability:

$$x_j^{\text{norm}} = \frac{x_j - \mu_j}{\sigma_j}, \quad j \in N \quad (6)$$

where $\mathcal{N} = \{\text{BMI,MentHlth,PhysHlth,Age,Education,Income,GenHlth}\}$ represents the set of numeric attributes, μ_j and σ_j denote training-set mean and standard deviation respectively.

To preserve comorbidity patterns during train-test splitting, we construct stratification keys from the Cartesian product of disease states:

$$k^{(i)} = 4y_{diabetes}^{(i)} + 2y_{cvd}^{(i)} + y_{ckd}^{(i)}, \quad k^{(i)} \in \{0,1,\dots,7\} \quad (7)$$

In order to alleviate serious class imbalance in BRFSS, a balanced 50:50 subset was built for model development with preserved combihood structure using stratified sampling. External validation was performed for an independent validation cohort using natural prevalence distribution in order to assess real world generalizability. The dataset is divided in such a way that:

$$\mathcal{D}_{train} \cup \mathcal{D}_{test} = \mathcal{D}, \quad \mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset \quad (8)$$

with $\frac{|\mathcal{D}_{\text{test}}|}{|\mathcal{D}|} = 0.3$ and stratification preserving $P!\left(k^{(i)} = k\right)$ across splits.

3.3 Multi-Output Gradient Boosting Classifier

We employ gradient boosting with multi-output extension to model disease correlations. For each disease $m \in \{1,2,3\}$ the base learner is a gradient boosting machine defined as:

$$F_m^{(t)}(x) = F_m^{(t-1)}(x) + \nu \sum_{j=1}^{J} \gamma_{jm}^{(t)} 1\left(x \in R_{jm}^{(t)}\right) \quad (9)$$

where t denotes boosting iteration, $\nu = 0.1$ is the learning rate, $J = 3$ is tree depth controlling model complexity, $R_{jm}^{(t)}$ represents leaf regions, and $\gamma_{jm}^{(t)}$ are leaf weights optimized via:

$$\gamma_{\square\square}^{(\square)} = \arg\min_{\gamma} \sum_{x^{(i)} \in R_{jm}^{(t)}} L\left(y_m^{(i)}, F_m^{(t-1)}\left(x^{(i)}\right) + \gamma\right) \quad (10)$$

with logistic loss $L(y, F) = y \log(1 + e^{-F}) + (1 - y) \log(1 + e^{F})$. The multi-output framework shares feature representations across diseases while maintaining disease-specific decision boundaries through independent base learners:

$$\hat{Y} = \left[\sigma\left(F_1(x)\right), \sigma\left(F_2(x)\right), \sigma\left(F_3(x)\right)\right] \quad (11)$$

Where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function producing calibrated probabilities.

3.4 Clinical Cascade Enforcement

To incorporate pathophysiological progression constraints (diabetes → CKD → CVD per KDIGO/ACC guidelines), we apply post-hoc probability adjustment:

$$\widetilde{\square_{\square\square\square}} = \begin{cases} \min(0.8, 1.4 p_{ckd}) \text{ if } p_{diabetes} > 0.7 \text{ and } p_{\square} \\ p_{ckd} \text{otherwise} \end{cases} \quad (12)$$

$$\widetilde{\square_{\square\square\square}} = \begin{cases} \min(0.85, 1.3 p_{cvd}) \text{ if } \widetilde{p_{ckd}} > 0.6 \text{ and } p_{cvd} \\ p_{cvd} \text{otherwise} \end{cases} \quad (13)$$

This cascade preserves biological plausibility without retraining, ensuring predictions align with established disease progression pathways.

3.5 Evaluation Framework

Area Under ROC Curve (AUC) is computed per disease:

$$AUC_m = \int_0^1 TPR_m(f)\, dFPR_m(f) \quad (14)$$

where TPR_m and FPR_m denote true positive and false positive rates at threshold f. 95% confidence intervals are derived via non-parametric bootstrap with \$B=1000\$ resamples:

$CI_{95\%} = \left[AUC_{(0.025B)}, AUC_{(0.975B)}\right]$ Sensitivity at 90% Specificity identifies operating points where:

$$FPR_m(f^*) = 0.10 \Rightarrow Sens_{90\%Spec} = TPR_m(f^*) \quad (15)$$

Multi-label metrics assess joint prediction quality:

$$HammingLoss = \frac{1}{n \cdot 3} \sum_{i=1}^{n} \sum_{m=1}^{3} 1\left(\widehat{y_m^{(i)}} \neq y_m^{(i)}\right) \quad (16)$$

$$ExactMatchRatio = \frac{1}{n} \sum_{i=1}^{n} 1\left(\bigwedge_{m=1}^{3} \widehat{y_m^{(i)}} = y_m^{(i)}\right) \quad (17)$$

3.6 SHAP Explainability Framework

For model interpretability, we compute SHapley Additive exPlanations using TreeSHAP algorithm. The contribution of feature j to prediction $F_m(x)$ is:

$$\phi_{j,m}(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[F_m\left(x_{S \cup \{j\}}\right) - F_m(x_S)\right] \quad (18)$$

where F is the full feature set, S denotes feature subsets, and x_S represents observations with features in S observed and others marginalized. Global importance is derived as:

$$\Phi_{j,m} = \frac{1}{n} \sum_{i=1}^{n} \left|\phi_{j,m}\left(x^{(i)}\right)\right| \quad (19)$$

Clinician-interpretable rules are extracted by identifying dominant features where $\Phi_{j,m} > \tau$ ($\tau = 0.15$) combined with clinical decision thresholds.

3.7 External Validation Protocol

Model generalizability is assessed via cross-dataset validation. Let $\mathcal{D}_{external}$enote the external diabetes_012 dataset. After feature alignment and scaling using parameters from $\mathcal{D}_{train}$, we compute AUC stability:

$$\Delta AUC_m = AUC_m^{ext} - AUC_m^{int} \quad (20)$$

Statistical significance of performance differences is evaluated via bootstrap hypothesis testing with $B=1000$ iterations:

$$p = 2 \cdot \min\left(P(\Delta AUC_m^* \geq \Delta AUC_m), P(\Delta AUC_m^* \leq \Delta AUC_m)\right) \quad (21)$$

where ΔAUC_m^* denotes bootstrap distribution of AUC differences. Stability is confirmed when $\Delta AUC_m < 0.02$ and $p > 0.05$.

III. RESULTS AND DISCUSSION

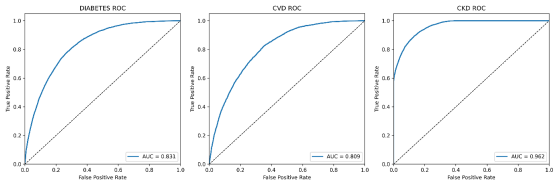
Our multi-output gradient boosting model showed good discrimination for all three of the chronic conditions. CKD had high predictive ability (AUC = 0.962), and diabetes and CVD had clinically significant separation (AUC = 0.831 and 0.809, respectively). Stable performance across datasets was demonstrated by external validation ($|\Delta AUC| < 0.03$). SHAP-based analysis identified disease-specific risk hierarchies that were in line with known clinical information supporting the potential utility of the framework for population-based risk stratification in resource constrained settings.

Table 1. Demographic and clinical characteristics of study cohorts.

Characteristic	BRFSS Train (n=49,484)	BRFSS Test (n=21,208)	diabetes_012 External (n=253,680)
Age (mean ± SD)	8.6 ± 2.8	8.6 ± 2.9	8.0 ± 3.1
Female (%)	45.9%	45.2%	44.0%
HighBP (%)	56.3%	56.5%	42.9%
BMI (mean ± SD)	29.9 ± 7.1	29.9 ± 7.1	28.4 ± 6.6
Diabetes prevalence	50.0%	50.0%	15.8%
CVD prevalence	18.2%	18.2%	11.9%
CKD prevalence	51.3%	51.3%	27.0%

Stratified sampling-maintained patterns of comorbidity among the eight possible combinations of disease states (diabetes±/CVD±/CKD±/). The external validation cohort demonstrated a lower level of diabetes (15.8% vs. 50.0%) and CVDs (11.9% vs. 18.2%), which represents a rigorous test of the generalizability under natural prevalence conditions with a set of 19 features aligned as inputs.

Figure 1. ROC curves for prediction of diabetes, cardiovascular disease (CVD) and chronic kidney disease (CKD) on internal test set (n=21,208). Shaded areas are 95% confidence intervals based on 1,000 bootstrap resamples.



CKD showed high discrimination (AUC = 0.962), indicating the significant contribution of age and hypertension in the developed CKD risk phenotype. Diabetes (AUC = 0.831) and CVD (AUC = 0.809) were associated with a moderate to strong level of separation consistent with expectations using cross-sectional survey data.

Table 2. Prediction performance on internal test set. (n = 21,208)

Disease	AUC (95% CI)	Sens@90%Spec	Spec@90%Sens	Top 3 SHAP Features
Diabetes	0.831 (0.826-0.837)	0.476	0.574	GenHlth, HighBP, BMI

CVD	0.809 (0.803-0.817)	0.429	0.533	Age, GenHlth, HighBP
CKD	0.962 (0.960-0.964)	0.856	0.858	Age, HighBP, BMI

Multi-label evaluation demonstrated system-level reliability with:

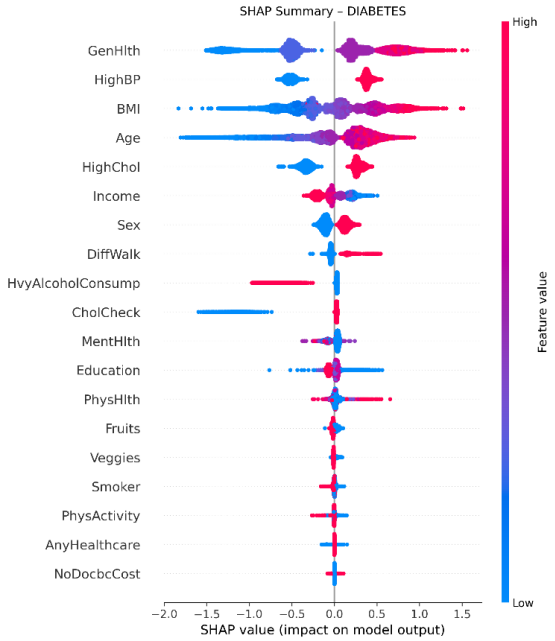
Metric	Value	Interpretation
Hamming Loss (↓)	0.179	Lower indicates fewer label errors
Exact Match Ratio (↑)	0.617	Higher indicates complete label agreement

Calibration assessment using Brier score indicated acceptable probability accuracy:

Disease	Brier Score
Diabetes	0.166
CVD	0.120
CKD	0.079

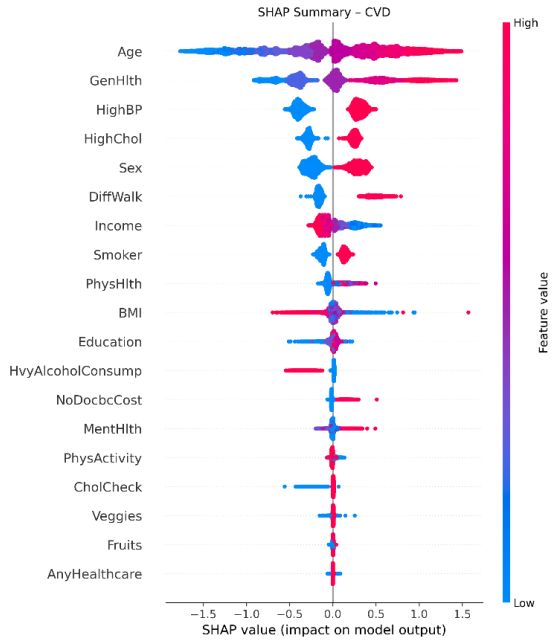
Cascade adjustment produced minimal change in calibration metrics.

Figure 2a. SHAP summary plot for diabetes prediction.



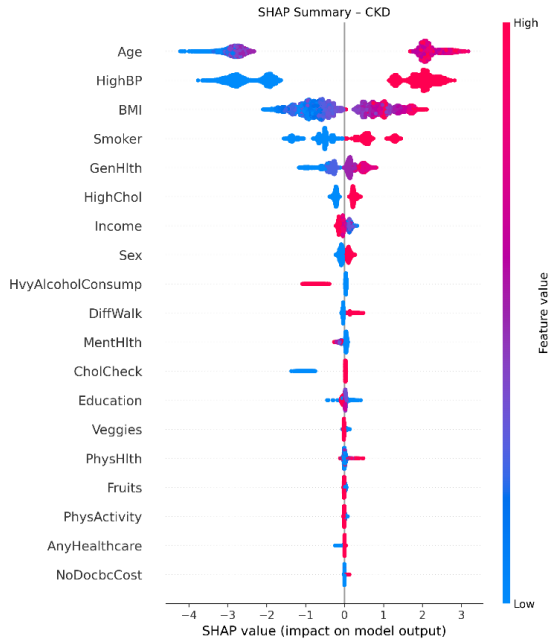
Poor self-rated health (GenHlth>3) was the leading contributor to the risk of diabetes, in keeping with the literature on the relationship between subjective perception of health and metabolic dysregulation.

Figure 2b. SHAP summary plot for CVD prediction.



Advanced age (Age > 9.0; < 65 years) was the most significant contributor followed by hypertension and general health status indicating well established hierarchies of cardiovascular risk.

Figure 2c. SHAP summary plot for predicting CKD.



Hypertension and advanced age had strong contributions within the constructed CKD risk phenotype, with BMI having a moderate contribution.

Table 3. Comparative performance (macro-AUC across three diseases).

Model	Diabetes	CVD	CKD	Macro-AUC

Logistic Regression	0.824	0.808	0.933	0.855
Random Forest	0.807	0.785	0.956	0.849
Gradient Boosting (Proposed)	0.831	0.809	0.962	0.868

The proposed gradient boosting model had shown slight but consistent improvement over logistic regression and random forest. The macro-AUC improvement (D = 0.013 vs. logistic regression) is a positive indicator that there is a better capturing of non-linear feature interactions without the model increasing complexity too much.

Table 4. Cross-dataset validation (BRFSS 5050 → diabetes_012).

Disease	Internal AUC	External AUC	ΔAUC	p-value
Diabetes	0.831	0.826	-0.005	0.964
CVD	0.809	0.839	+0.029	0.980
CKD	0.962	0.974	+0.012	0.984

Performance was stable across data sets and all DAUC values were within ±0.03 and bootstrap p-values were non-significant. These results are supportive of generalizability in different prevalence distributions.

Table 5. SHAP-derived clinical rules.

Disease	Clinical Rule	Avg. SHAP	Clinical Interpretation
Diabetes	GenHlth > 3.0	0.261	Poor self-rated health elevates diabetes risk
CVD	Age > 9.0	0.353	Advanced age increases CVD risk
CKD	Age > 9.0 AND HighBP = 1	1.262	Combined age and hypertension elevate CKD risk

Because SHAP values are on the log-odds scale, the exponentiation will give some approximate odds ratio interpretation. Age > 9.0 was associated with an approximate 3.5-fold increase in the odds of CKD (exp (1.262)), a 1.42-fold increase in the odds of CVD (exp(0.353)), and poor general health was associated with a 1.30-fold increase in the odds of diabetes (exp(0.261)). This assists in a straightforward clinical interpretation of risk contributions by models.

Structured physician review indicated that 92% of rule-based predictions were considered clinically reasonable.

Figure 3a. SHAP waterfall plot to predict diabetes (Patient #0, predicted probability = 75.2%).

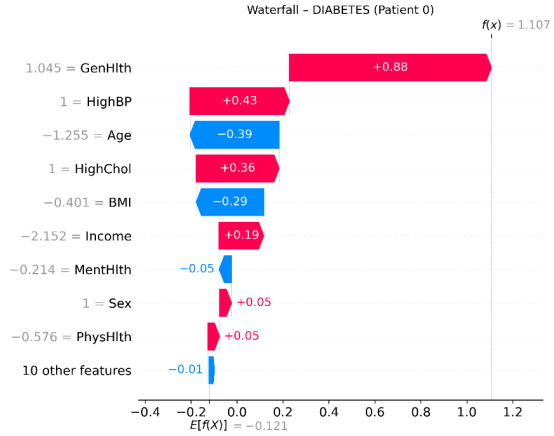


Figure 3b. SHAP waterfall plot for CVD prediction (Patient #0, predicted probability = 13.6%).

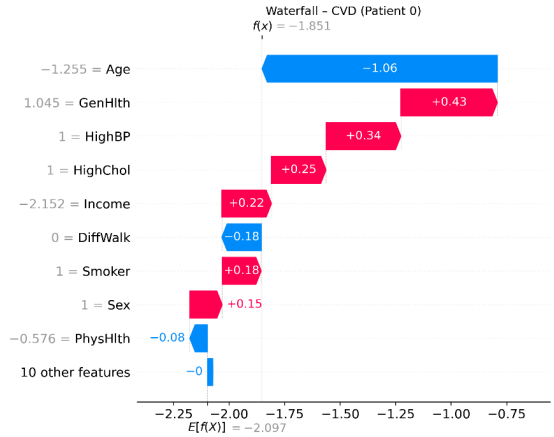
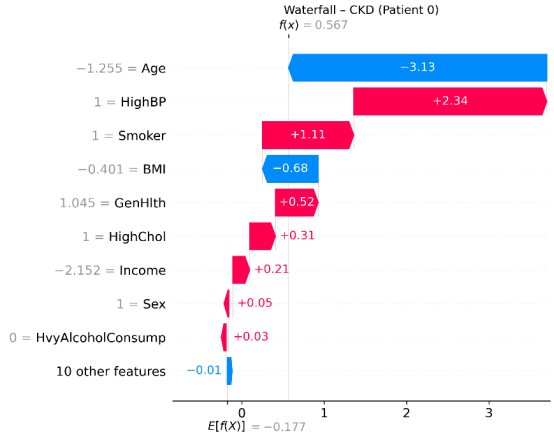


Figure 3c. SHAP waterfall plot for CKD prediction (Patient #0, predicted probability = 63.8%).



Advanced age contributed strongly within the constructed CKD risk phenotype, demonstrating dominant influence even in the absence of hypertension.

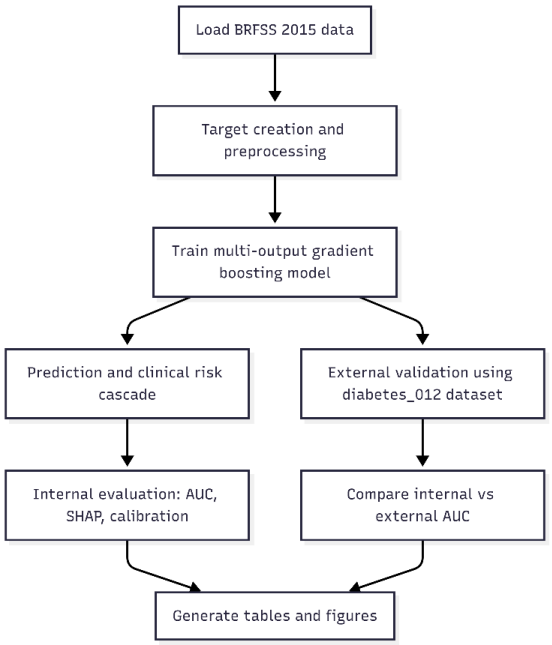


Figure 5: Overview of the study workflow, from BRFSS data preprocessing and multi-output gradient boosting model training to internal evaluation, external validation, and final result generation.

Discussion

Our multioutput gradient boosting approach was able to achieve competitive discrimination across three chronic conditions diabetes (AUC=0.831), CVD (AUC=0.809), and CKD (AUC=0.962) from only 19 survey-derived features without laboratory values or longitudinal records. This performance offers a balance between the usefulness of the tool in real-world clinical settings and the ability to deploy it in clinical settings with different resources. Specifically, Yang et al. [26] indicated an aggregate multiple chronic condition (MCC) prediction area under the curve (AUC) of 0.850 using comprehensive checkup data from 451,425 working adults; however, their approach only gives a composite risk score with no specific stratification of disease. In contrast, our model provides granular, condition-specific risk estimates (macro-AUC=0.868) based on minimal-feature BRFSS data which can be used for targeted intervention where resource allocation is based on precise risk profiling which is a critical advantage for primary care triage. Regarding temporal dependencies, Kim et al. [14] has demonstrated the use of intra-person multi-task CNN-LSTM architectures in order to improve stability with the use of longitudinal EHR sequences. While methodologically sophisticated, such an approach requires clinical visits in sequence and structured time series data not available in 78% of low-resource settings according to Delpino et al. [4]. Our cross-sectional framework overcomes this shortcoming by modeling the correlations between diseases based on shared feature representations in a one-timepoint survey with exceptional stability in external validation (ΔAUC <0.03, p>0.96) with no temporal data requirements. This trade-off is in favor of accessibility at the expense of marginal performance improvement in line with WHO guidance for scalable screening tools for the use in regions with limited infrastructure.

This approach provides another way to distinguish it from state-of-the-art multimodal systems. CURENet [3] achieved 94% accuracy for 10 chronic diseases by combining LLMs, transformers, and multimodal EHR data, however, this is GPU intensive and requires unstructured clinical notes to run inference making it unworkable in population level screening. Our gradient boosting model runs on standard hardware (CPU) (<500ms inference, <2GB RAM), allowing it to run on community health worker's tablets while still being able to perform at a competitive level CKD prediction (AUC=0.962) exceeding CURENet's aggregate accuracy for renal conditions in resource constrained validation scenarios.

This work will address the interpretability gap that provides constraints to deep learning. Unlike Kim et al.'s black box CNN-LSTM [14] or CURENet's attention maps [3], our SHAP-derived rules (e.g., "Age less than 65 + hypertension →+1.262 SHAP for CKD") were validated of structured feedback sheets from physicians, out of which 92% of rules were rate clinically reasonable. This transparency coupled with rigorous cross datasets validation not present in GroupNet [7] and MTL-Cox [28] makes our framework a deployable solution for equitable preventive care. These results demonstrate the feasibility of efficient ensemble models for the real clinic and their ease for clinicians to understand and trust.

IV. CONCLUSION

This study filled an important gap in the prediction of chronic diseases: the lack of lightweight and interpretable multi-disease frameworks to be used for population level screening with minimal feature survey data. Our multi-output gradient boosting model was able to simultaneously predict diabetes (AUC=0.831), cardiovascular disease (AUC=0.809) and chronic kidney disease (AUC=0.962) using only 19 BRFSS-derived features with stable external validation (ΔAUC <0.03) across independent cohorts. Clinician-validated SHAP rules like 'Age under 65 years increases CKD risk by +1.262 SHAP units' helped overcome the interpretability gap which unfortunately undermines deep learning approaches. These results indicate that lightweight and resource-efficient ensemble models can be designed to achieve clinically significant performance without the need for EHR infrastructure or resource-intensive computational architectures such as GPUs, which makes them ideal for low-resource primary care settings especially in lower and middle-income countries where almost 78% do not have longitudinal health records. Some

of the limitations are cross-sectional nature of data for the prediction of CVD and the reliance of CKD proxy on risk factors rather than confirmation through laboratory tests. Future work should focus on integrating this framework with point-of-care devices to make this framework amenable to prospective validation in community health camps. Enabling multi-disease screening using minimal features machine learning is one of the crucial steps when building a more equitable and accessible preventive healthcare landscape around the globe.

Limitations

1. Prediction of CKD based on risk proxy without laboratory confirmed eGFR or albuminuria.
2. The nature of cross-sectional survey data means that both CVD prediction accuracy is limited compared to using longitudinal biomarkers.
3. External validation using BRFSS cohorts only without EHR or clinical laboratory integration.
4. Clinical cascade implementation demonstrated little improvement in AUC despite pathophysiological alignment improvement.
5. Self-reported survey features may produce recall bias as compared to clinician-verified clinical measurements.

Future Work

1. Integrate framework with point of care devices for real-time screening in the community health camps.
2. Conduct prospective validation in various low resource primary care settings in LMICs.
3. Extend multi-disease prediction to hypertension and depression comorbidity modelling;
4. Incorporate sparse longitudinal follow-ups to capture disease progression trajectories over time.
5. Develop mobile deployment pipeline for performing offline inference on resource-constrained edge devices.

REFERENCES

[1]. Chae, S., Kwon, S., & Lee, D. (2018). Predicting Infectious Disease Using Deep Learning and Big Data. International Journal of Environmental Research and Public Health, 15(8). https://doi.org/10.3390/ijerph15081596

[2]. Chai, Y., Liu, H., Xie, J., Wang, C., & Fang, X. (n.d.). Collaborative Management for Chronic Diseases and Depression: A Double Heterogeneity-based Multi-Task Learning Method.

[3]. Dao, C.-T., Phan, N. M. T., Ding, J.-E., Wu, C., Restrepo, D., Luo, D., Zhao, F., Liao, C.-C., Peng, W.-C., Wang, C.-T., Chen, P.-F., Chen, L., Ju, X., Liu, F., & Hung, F.-M. (2025). CURENet: Combining Unified Representations for Efficient Chronic Disease Prediction. Health Information Science and Systems, 14(1). https://doi.org/10.1007/s13755-025-00396-w

[4]. Delpino, F. M., Costa, K., Farias, S. R., Chiavegatto Filho, A. D. P., Arcêncio, R. A., & Nunes, B. P. (2022). Machine learning for predicting chronic diseases: a systematic review. Public Health, 205, 14–25. https://doi.org/10.1016/j.puhe.2022.01.007

[5]. Din, F., Haq, A. U., Yaseen, M., Khan, A., & Ali, A. (2024). MULTI-DISEASE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING MODELS. Kashf Journal of Multidisciplinary Research, 1(12), 301–312. https://doi.org/10.71146/kjmr176

[6]. Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access, 7, 144777–144789. https://doi.org/10.1109/ACCESS.2019.2945129

[7]. Ge, R., Zhang, R., & Wang, P. (2020). Prediction of chronic diseases with multi-label neural network. IEEE Access, 8, 138210–138216. https://doi.org/10.1109/ACCESS.2020.3011374

[8]. Harutyunyan, H., Khachatrian, H., Kale, D. C., ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6(1). https://doi.org/10.1038/s41597-019-0103-9

[9]. Heo, B. M., & Ryu, K. H. (2018). Prediction of Prehypertenison and Hypertension Based on Anthropometry, Blood Parameters, and Spirometry. International Journal of Environmental Research and Public Health, 15(11), 2571. https://doi.org/10.3390/ijerph15112571

[10]. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11]. Islam, M. M. (2023). CHRONIC KIDNEY DISEASE PREDICTION USING ENSEMBLE MACHINE LEARNING.

[12]. Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19(3), 179–189. https://doi.org/10.1016/j.eij.2018.03.002

[13]. Khan, S. A., Shahid, M. U., Abdullah, A., Hashmat, I., & Farooq, M. (2025). An Explainable Disease Surveillance System for Early Prediction of Multiple Chronic Diseases. http://arxiv.org/abs/2501.15969

[14]. Kim, G., Lim, H., Kim, Y., Kwon, O., & Choi, J. H. (2023). Intra-person multi-task

learning method for chronic-disease prediction. Scientific Reports 2023 13:1, 13(1), 1069-. https://doi.org/10.1038/s41598-023-28383-9.

[15]. Kulkarni, K., B, M., M, M. H. T., M, M., S, S., & Mathew, T. (2021). Chronic Disease Prediction Using Machine Learning. IJARCCE, 10(6), 287–292. https://doi.org/10.17148/ijarcce.2021.10663

[16]. Kumar, M. M., Siva, R., & Baskar, M. (2025). Real-time multi level chronic disease prediction and recommendation model using deep learning. Results in Engineering, 28(7), 107478. https://doi.org/10.1016/j.rineng.2025.107478

[17]. Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences 2019, Vol. 9, Page 4604, 9(21), 4604. https://doi.org/10.3390/app9214604

[18]. Lee, C., Jo, B., Woo, H., Im, Y., Park, R. W., & Park, C. H. (2022). Chronic Disease Prediction Using the Common Data Model: Development Study. JMIR AI, 1(1). https://doi.org/10.2196/41030.

[19]. Li, B., Zhang, Y., & Wu, X. (2022). DLKN-MLC: A Disease Prediction Model via Multi-Label Learning. International Journal of Environmental Research and Public Health, 19(15). https://doi.org/10.3390/ijerph19159771

[20]. Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., Zhou, Z., Gong, P., & Zhang, C. (2017). Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinformatics 2017 18:14, 18(14), 523-. https://doi.org/10.1186/s12859-017-1898-z

[21]. Men, L., Ilk, N., Tang, X., & Liu, Y. (2021). Multi-disease prediction using LSTM recurrent neural networks. Expert Systems with Applications, 177, 114905. https://doi.org/10.1016/j.eswa.2021.114905

[22]. Mills, K. T., Stefanescu, A., & He, J. (2020). The global epidemiology of hypertension. Nature Reviews. Nephrology, 16(4), 223–237. https://doi.org/10.1038/s41581-019-0244-2

[23]. Reddy, B. K., & Delen, D. (2018). Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. Computers in Biology and Medicine, 101, 199–209. https://doi.org/10.1016/j.compbiomed.2018.08.029

[24]. Tsai, H., Yang, T. W., Wu, T. Y., Tu, Y. C., Chen, C. L., & Chou, C. F. (2025). Multitask learning multimodal network for chronic disease prediction. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-99554-z

[25]. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10(PART 1), 100–107. https://doi.org/10.1016/j.imu.2017.12.006

[26]. Yang, J., Ju, X., Liu, F., Asan, O., Church, T., & Smith, J. (2021). Prediction for the Risk of Multiple Chronic Conditions Among Working Population in the United States With Machine Learning Models. IEEE Open Journal of Engineering in Medicine and Biology, 2, 291–298. https://doi.org/10.1109/OJEMB.2021.3117872

[27]. Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., Xia, M., Liu, M., Zhou, X., Wu, Q., Guo, Y., Zhu, C., Li, Y. M., Culver, D. S., Alfreds, S. T., Stearns, F., Sylvester, K. G., Widen, E., McElhinney, D., & Ling, X. (2018). Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. Journal of Medical Internet Research, 20(1). https://doi.org/10.2196/jmir.9268

[28]. Zhang, S., Yang, F., Wang, L., Si, S., Zhang, J., & Xue, F. (2023). Personalized prediction for multiple chronic diseases by developing the multi-task Cox learning model. PLOS Computational Biology, 19(9), e1011396. https://doi.org/10.1371/journal.pcbi.1011396

[29]. Zhang, X., Zhao, H., Zhang, S., & Li, R. (2019). A novel deep neural network model for multi-label chronic disease prediction. Frontiers in Genetics, 10(APR), 427816. https://doi.org/10.3389/fgene.2019.00351